Watchtower Peer Review

Data Science Division, Data Ethics Secretariat

Statistics Canada

Introduction

This document provides a peer review of the Watchtower Risk Identification Prototype. The core purpose is to inform the team responsible for this prototype of potential ethical or quality concerns. The review has been done by relying on Statistics Canada (2021) Responsible Machine Learning Framework. Bronson and Millar (2020) propose a high level process for peer review under Canada's Directive on Automated Decision-Making which was also consulted. Moreover, we have examined the Algorithmic Impact Assessment form as it has been filled by the team responsible for the Watchtower project.

Major Recommendations

1. We recommend describing, with concrete examples, the kind of decisions/actions that the IRCC officials will be able to make/take with this tool in order for the readers to better understand the benefits and the risks associated with this project. What will the agents do with the information? What can happen to the applicants? Are the applicants in Canada or in a different country? It would be good to define the expression 'adverse information' more rigorously. Is it an euphemism of 'crime' or 'offense'?
2. The algorithm associates "data patterns to adverse information" p.5. We recommend using more standard terminology such as "independent variables to dependent variables" or "features to labels". This will improve the interpretability of the documentation of the project.
3. Based on the information in the document, this looks like a predictive machine learning project where one models the relationships between the variables in the dataset. How can this tool provide "fact-based information that end users (IRCC officers) may find relevant for their adjudication process" p.5 if the officers are not going to classify/make predictions on individuals? Expressions such as "identify only those patterns suggestive of organized risk" p.8 imply predictive labelling. Therefore, we recommend removing expressions such as "the tool is not predictive"p.5.
4. If this is a predictive modelling project, then a train/validation/test methodology must be applied. If there is no need for validation, then it should be explained. The metrics that will be used should be described and justified. How do you measure the "utility and timeliness" p.14 ?

5. We recommend a more thorough description of the quality of the data set that will be used by the algorithm. Does the data fail to cover a subpopulation that could be at risk? Are there any missing data? If so, has there been any kind of imputation (without leakage)? Do you use all of the available data to run the algorithm or only the observations that have been previously flagged by officers?
6. We recommend answering the following two questions. Will the information used reflect and exacerbates stereotypes? We understand that the purpose is not to identify "broad subpopulation that tends to have higher adverse rates" but will someone monitor this over time?
7. We recommend describing how the quality of the algorithm will be maintained over time. How will you measure the "utility and timeliness" p.14 over time? Will an individual be investigated over and over again even if the first investigation did not find anything suspicious?
8. It would be good to describe the alternative methods that have been considered, like decision trees, and to explain why they have not been chosen.
9. We recommend that this report should be approved by the Deputy Minister in charge of the program that will use the automated decision system.
10. We recommend the publication of this report and the completed Algorithmic Impact Assessment on the Open Government Portal prior to deploying the system.

Minor Recommendations

1. We recommend explaining how this tool will ultimately benefit Canadians in terms that they will understand. This can help motivate the need for such measures.
2. We recommend indicating that some research has been done (bibliography) on the ethical use of machine learning for decision making with crime data.

Reference

1. Statistics Canada (2021). "Framework for Responsible Marching Learning Processes at Statistics Canada", https://www150.statcan.gc.ca/n1/pub/89-20-0006/892000062021001-eng.htm
2. Bronson and Millar (2020). "Peer Review for Automated Decision-Making Tools Under Canada's Directive on Automated Decision-Making", internal report based on a study with Treasury Board Secretariat and Canadian School of Public Service

NRC·CNRC Data Analytics Centre

Temporary Resident applications Volume Management:

NRC Evaluation

Stéphane Tremblay

July 12th, 2018

National Research Council Canada | Conseil national de recherches Canada

Canada

Contents

Executive Summary

La méthodologie de l'initiative est excellente et suit les étapes nécessaires au succès d'un projet d'apprentissage machine. L'initiative est très bien adaptée aux différents risques organisationnels (légales, perception du public, sécurité) tout en maximisant les mesures de performances, c'est-à-dire la transparence de la méthodologie, l'efficacité des opérations et la qualité des données.

L'analyse fourni 9 recommandations et 3 projets de collaborations avec le CNRC. Les sujets abordés vont des modèles d'apprentissage automatique, de l'environnement de travail, des algorithmes ainsi que de reproductibilité des résultats.

1. Objectives

Immigration, Refugees and Citizenship Canada (IRCC) requires the services of NRC to conduct a scientific/methodological review of the tools and approaches that are used for the deployment of predictive models that will be use as a solution to support TRVM initiative.

The rational for conducting for the undertaking of a scientific/methodological review by a third independent party is to ensure both the integrity, robustness and quality of the work that had been accomplished so far.

In principle, a cursory evaluation of key phases such as Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment. Since the project is relatively nascent the focus will be restricted as described in the "Scope of Work" below.

Once the cursory review is completed, both participants will determine whether a more detailed evaluation would be required.

2. Background

The Advanced Analytics Lab (AAL) is responsible to undertake the TRVM-AA project to support the Operations Sector in achieving client service and integrity risk management outcomes in several lines of business described below.

At a high level, the approved scope of the project includes:

1. Deployment of predictive models for:
   - India and China Temporary Resident Visa applicants (eApp only)
   - India and China Study Permits (eApp only)
   - Citizenship Risk Triage
   - Passport Risk Triage
2. Development of forecasting models
3. Development of a Business Intelligence and Exploratory Environment

3. Requirement

Scientific/methodological review is essentially a "peer" review process of applying expert knowledge of acceptable criteria to determine whether the tools and protocol are adequate.

A high level report will describes what was accomplished so far including an assessment of the overall merits and identification of the tools and protocols that have been put in place.

Scientific/methodological review is a constructive process. The intent of the review is to assist IRCC to meet a minimally adequate set of criteria for the work that have been accomplished.

4. Scope of Work

The requirement is to engage the services of NRC to supply professional services expertise on the following objectives:

1. Review and assess key phases of the work undertaken so far in the following areas :
    a. Data understanding in particular on data quality;
    b. Data preparation: data cleaning technics and results, derived attributes and generated records, integration data approach and aggregations;
    c. Modelling: modelling technique and assumptions, test design, build model (settings, models, descriptions), assessment and revise parameter settings;
    d. Evaluation (efficiency of the model, deficiencies).

5. Tasks

The tasks required of NRC will include the following:

1. Develop action plan outline to conduct the review
    a. Review documentation
    b. Meet key stakeholders
2. Deliverables and Acceptance Criteria:
    a. The Plan
    b. An interim Report
    c. Final Report on the detailed and comprehensive review of the method.

6. Activities

1. Two one-day meetings took place at Immigration, Refugees and Citizenship Canada on March 20th and 27th respectively.
2. A plan and an interim reports were sent following each of those meetings
3. This is the final report

7. Analysis

La méthodologie de l'initiative est excellente et suit les étapes nécessaires au succès d'un projet d'apprentissage machine. L'initiative est très bien adaptée aux différents risques organisationnels (légales, perception du public, sécurité) tout en maximisant les mesures de performances, c'est-à-dire la transparence de la méthodologie, l'efficacité des opérations et la qualité des données.

L'analyse a également vérifiée les modèles d'apprentissage automatique, l'environnement de travail, les algorithmes utilisés afin de fournir des recommandations. Une attention particulière a été mise sur la reproductibilité des résultats.

Avec l'exception d'une activité redondante de surajustement (over-fitting), l'environnement, l'approche et les algorithmes utilisés semblent être adéquats pour répondre au besoin du projet et sa mise en production. Dans les itérations futures, une attention plus équilibrée de la variabilité du modèle et le biais serait bénéfique. L'approche utilisée dans son ensemble pour garantir la reproductibilité des résultats est excellente, simple et claire.

7.1. Environnement

L'environnement SPSS est utilisé pour tout le traitement, la modélisation et le déploiement. Cet environnement est stable et bien supporté. Son désavantage est son manque de flexibilité; les options sont limitées à ce que SPSS offre.

7.2. Cadre

L'enjeu est de classifier des demandes de visas pour la Chine. Il y a 3 ans de données étiquetées (approuvée ou non). L'entraînement des données se fait à l'écart (batch) et la performance est mesurée à l'aide de la précision de la prédiction des visas approuvés. L'approche utilisée suit les normes utilisées dans le domaine de l'apprentissage automatique.

7.3. Petite population (débalancement)

Le nombre de Visa refusé est relativement petit par rapport à celui des Visas approuvés. Les techniques de sous-échantillonnage utilisées pour réduire cet enjeu sont excellentes. Étant donné son impact réducteur sur la taille de l'échantillon, il serait utile de mieux comprendre l'effet de ce débalancement. Seulement une approche a été testée.

7.4. Modélisation

La modélisation utilise 3 années de données. La partie entraînement est composée d'applications des 3 années. La partie cross-validation (appelé testage) est composée d'application de la dernière année seulement. Et la partie testage (appelé validation) est composée de mois ne faisant ni parti entraînement ou cross-validation.

La technique de modélisation utilisée pour ce projet est un arbre décisionnel (decision tree) qui est un excellent outil reconnu pour sa transparence mais relativement moins précis que d'autres. En revanche quelques techniques de modélisation pourraient être améliorées et d'autres annulées. Une meilleure représentativité de la partie cross-validation à celle d'entraînement pourrait donner de meilleurs résultats. L'utilisation de chiffres aléatoires (seed) pour déterminer le meilleur modèle est inutile et met une pression inutile sur le over-fiting. Il semble que les données plus vieilles ont moins de valeurs que les plus récentes. Ces deux enjeux montrent que les actions liées à la modélisation sont de nature à réduire le biais au profit de la variance du modèle.

7.5. Réentraînement

La décision de réentraîner les modèles tous les trois mois est excellente.

7.6. Reproductibilité

L'utilisation du logiciel SPSS, la documentation et la justification de chaque variable, l'utilisation de « seed », la transparence du modèle choisi sont excellents pour une reproductibilité des résultats et un transfert des connaissances. Le risque d'enjeux liés à la reproductibilité est réduit à son minimum.

8. Recommendations

Suite aux deux jours de consultations, neuf recommandations ont été identifiées.

1. Avoir une stratégie de surveillance sur le réapprentissage et mise à jour des paramètres au besoin
2. Modifier l'unité d'analyse à « famille » au lieu de « personne »
3. Uniformiser l'apprentissage au groupe « Validation » (aussi appelé Testing dans la littérature) pour que les données entraîner soit semblables aux données tests.
4. Dans un environnement parallèle, développer des modèles plus sophistiqué en commençant par : Random Forest, SVM, Neural network, Deep learning, etc.
5. Augmenter, voir doubler ou tripler, la taille du groupe training (approuvé) même si le nombre de refus est relativement petit.
6. Augmenter l'échantillon en combinant les données de d'autres pays d'origine.

7. Utilisation du « clustering » pour mieux comprendre les différences entre les saisons, les pays, etc. et l'analyse de la fraude.
8. Considérer l'utilisation de logiciels libres tels R ou Python pour expérimenter de nouvelles approches et avoir plus de flexibilité sur la modélisation
9. Expérimenter les différentes formes de débalancement. Identifier l'approche qui optimise la précision des résultats.

9. Future Collaborations

Différentes opportunités de collaboration entre le IRCC et le CNRC ont également été abordées:

1. Saisir l'information capturée sur des images et l'apparier à une base de données (passeport et visa)
2. Utiliser les plateformes de calculs du CNRC pour tester des modèles plus complexes.
3. Développer du codes pour des modèles plus sophistiqués (en R ou Python par exemple)

SCLPC Peer Review

Data Science Division, Data Ethics Secretariat

Statistics Canada

Introduction

This document provides a peer review of the Spouse or Common-Law Partner in Canada Class (SCLPC) Advanced Analytics project based on the documentation received from the SCLPC team. The core purpose is to inform the team responsible for this prototype of potential ethical and quality concerns. The review has been done by relying on Statistics Canada Responsible Machine Learning Framework. The document entitled "Peer Review for Automated Decision-Making Tools Under Canada's Directive on Automated Decision-Making", has also been consulted. Moreover, we have examined the Algorithmic Impact Assessment form as it has been filled by the team responsible for the SCLPC project.

Acknowledgement of best practices implemented by this project

1. An accountability governance model where the Director General of the branch is accountable for the modelling results and their use.
2. Steps taken for model explainability and rules justification reviewed by subject matter experts. Further comments in the recommendations section.
3. A Model Privacy Assessment was completed.
4. A quality assurance plan is in place to monitor the model performance over time.
5. An overview of how advanced analytics is being used in this project has been published in the IRCC webpage for transparency.
6. A gender-based analysis plus (GBA+) was performed for preventing from discrimination.

Major Recommendations

1. A supervised machine learning workflow usually involves training, validation, and test set. There is no mention of a validation set in the documentation. Usually a validation set is used to optimize hyperparameters of the method. Is there any hyperparameters used by the proposed method? If yes, we recommend explaining why and answer any potential leakage concerns that one might have. For example, it would be inadequate to make changes (manually or otherwise) to the final model/decision rules, based on the quality metrics computed with the test set.

2. We recommend addressing the issue of the independence of the observations. Are the training and test set similar to the data that will be obtained in the future? Are there any time/historical dependencies? For example, did the context of the pandemic affect the content of the data in such a way that future data might be very different? If so, statistics computed from the test dataset might be biased.
3. What is the negative impact of wrongly placing a case in the Green Bin? How many individuals could be affected by this mistake? If there is no significant negative impact, then why do we need an interpretable model since the cases that are not placed in the Green Bin will be reviewed by an agent anyway?
4. Bias detection and mitigation
   4.1. The document mentions efforts to reduce bias without giving further details. We recommend explaining the type of bias that was considered and examined for this project and how it was mitigated.
   4.2. It was not clear if the analytical dataset used for analysis was a sample or a census (whole population).
   If a sample,
   - What was the sampling design? Stratification, other? Examples of (good stratification for) representation: time (2018, 2019), treatment centers, programs, other.
   - What was the breakdown of the analytical dataset over 2018 and 2019? Was it 50/50? Other? (We are interested in knowing if there has been a good representativeness over time)
   - In terms of Train/Test, were the proportions 70/30 for both 2018 and 2019?
   - Did we assume that both 2018 and 2019 populations are similar?

   4.3. In terms of bias, what is the impact of the post manual adjustments/ tweaks to model rules?
   4.4. In page 8, we read that the "model" was built using a dataset of 39,190 applications and then tested on the same dataset to come up with a precision of 99.3%. So here, the performance was not assessed on the TEST set but on the whole data (TRAIN + TEST)? The performance on TRAIN is usually (relatively) high compared to TEST.
5. The document mentions that privacy principles have been taken into considerations (i.e. data minimization, reducing data granularity, de-identification and need-to-know) but it would be good to add some concrete examples that show how this has been implemented.
6. We recommend adding some more quality metrics (confusion matrix for example) and how the challenge of class imbalances has been addressed. Similarly, it would be good to add a few words on the quality of the data and the challenges that it provided (e.g. missing data).
7. To assess the technical part of this work, it would have been nice to have access to a technical report with more details on the data, data processing, modelling, and all the assumptions made.

Minor Recommendation

1. We recommend explaining how this tool will ultimately benefit Canadians in terms that they will understand. By answering the question "Why is it important to reduce the processing time for the applicants?" we can add to the value of the project.